Medida de similitud basada en saliencia

Sergio Domínguez

Resumen

La proliferación en todos los ámbitos de la producción multimedia está dando lugar a la aparición de nuevos paradigmas de recuperación de información visual. Dentro de éstos, uno de los más significativos es el de los sistemas de recuperación de información visual, VIRS (*Visual Information Retrieval Systems*), en los que una de las tareas más representativas es la ordenación de una población de imágenes según su similitud con un ejemplo dado. En este trabajo se presenta una propuesta original para la evaluación de la similitud entre dos imágenes, basándose en la extensión del concepto de saliencia desde el espacio de imágenes al de características para establecer la relevancia de cada componente de dicho vector. Para ello se introducen metodologías para la cuantificación de la saliencia de valores individuales de características, para la combinación de estas cuantificaciones en procesos de comparación entre dos imágenes, y para, finalmente, establecer la mencionada ponderación de cada característica en atención a esta combinación. Se presentan igualmente los resultados de evaluar esta propuesta en una tarea de recuperación de imágenes por contenido en comparación con los obtenidos con la distancia euclídea. Esta comparación se realiza mediante la evaluación de ambos resultados por voluntarios.

Palabras Clave:
Bases de datos de imágenes, recuperación basada en contenido, medidas de similitud, modelos perceptuales, análisis de imágenes

1. Introducción

La enorme cantidad de material audiovisual accesible en la actualidad plantea el problema de su recuperación, con lo que surgen los sistemas de recuperación de información visual (VIRS, *Visual Information Retrieval Systems*). Dentro de este marco, las dos tareas que típicamente se pueden plantear son el *matching* y la recuperación por similitud. El primero hace usualmente referencia a la tarea de recuperar una instancia de un objeto cuya apariencia puede verse modificada por distintos motivos (cambios de color, iluminación, oclusiones, ...) (Santini and Jain, 1999). En cuanto al segundo, el problema se plantea en términos de recuperar los elementos de una base de datos que se caractericen por ser parecidos al objeto presentado como *query*. En muchos casos, la forma de caracterizar las imágenes y las técnicas utilizadas en la resolución de ambas tareas son similares; se parte de una representación de un determinado estímulo visual mediante la elección de una serie de descriptores, con lo que éste se puede modelar como un punto dentro de un espacio de características. A continuación, se define una función que realiza una comparación entre los distintos estímulos mediante la evaluación de la proximidad de los puntos correspondientes a cada uno de ellos en dicho espacio.

En el caso del matching, el éxito o el fracaso de la aplicación se cifra en su capacidad de recuperar dichas instancias, que básicamente representan al mismo objeto. Sin embargo, en el segundo, el éxito se obtendrá en la medida en la que el resultado presentado sea convincente para el usuario que ha lanzado la búsqueda: en este sentido, dicho usuario deberá percibir las imágenes recuperadas como similares a la utilizada como consulta. Lógicamente, en la consecución de este objetivo debe intervenir un modelo de evaluación de similitud que refleje la percepción de parecido por parte del usuario.

El estudio de la percepción humana de similitud tradicionalmente se ha llevado a cabo en el ámbito de la Psicología, habiéndose alcanzado resultados muy significativos que han llegado a explicar, hasta cierto punto, este comportamiento. Fruto de estos resultados se han creado distintos modelos de evaluación de similitud, aplicados con mayor o menor éxito en distintos sistemas de recuperación de imágenes por contenido.

Dentro de los diferentes estudios psicológicos realizados con objeto de caracterizar la percepción se similitud, uno de los más influyentes es sin duda el descrito en (Tversky, 1977) y (Tversky and Gati, 1982). En él se establece un claro vínculo entre la percepción de similitud y la saliencia de alguna o va-

rias características de los estímulos comparados. Fruto de esta correlación, Tversky propuso su famoso *modelo de contraste* para la evaluación de similitud entre estímulos representados por características cualitativas.

Siguiendo con esta línea de trabajo que relaciona similitud y saliencia, en este artículo se propone un procedimiento para establecer una medida de similitud entre dos estímulos visuales basada en la utilización de ésta como clave perceptual. Para este fin se propone, en primer lugar, una metodología para la cuantificación de la saliencia de una característica; a continuación, y basándose en esta cuantificación obtenida en un par de imágenes, un procedimiento para evaluar la relevancia de una característica durante un proceso de medida de similitud; finalmente, y sobre la base de la relevancia calculada, un método para establecer la ponderación adecuada a la comparación de los valores de dicha característica medida sobre ambas imágenes.

En la Sección 2 de este artículo se realiza un repaso de los hallazgos más significativos relacionados con los conceptos de saliencia y similitud, enmarcados fundamentalmente en el estudio de la atención visual, así como de otros trabajos en la misma dirección que el que ahora se presenta.

En la Sección 3 se presenta el modelo de evaluación de similitud propuesto, que profundiza en la relación existente entre la saliencia de una o varias caracterísitcas de un estímulo y el resultado de su comparación con otro; para ello se introducen las hipótesis fundamentales, que se apoyan en los desarrollos teóricos y experimentales expuestos en la Sección 2, y que se unifican en la función de evaluación de similitud presentada en este trabajo.

En la Sección 4 se presentan los resultados experimentales alcanzados con el modelo propuesto; éste fue sometido a una tarea de evaluación de similitud, y los resultados alcanzados se cotejaron con los obtenidos de una serie de voluntarios sometidos a la misma tarea. La comparación entre la evaluación del modelo propuesto y de observadores humanos es igualmente presentada. Finalmente, la Sección 5 se dedica al establecimiento de las pertinenetes conclusiones.

2. El modelo perceptual: similitud y saliencia visual

En esta sección se van a describir los estudios más relevantes publicados hasta la fecha en los campos de caracterización de la similitud entre dos estímulos, así como sobre la saliencia como mecanismo clave para la comprensión de la resolución de tareas visuales.

2.1. Relevancia visual: saliencia

El concepto de saliencia se utiliza para explicar los procesos que rigen la atención visual en humanos y primates. La teoría psicológica comúnmente aceptada afirma que la atención se materializa mediante dos procesos que se suceden; en primer lugar, uno preatentivo, en el que la información recogida por el sistema visual se procesa muy rápidamente y en paralelo, y cuyo resultado viene dado exclusivamente por la configuración del propio estímulo; en segundo lugar otro atentivo, que se realiza secuencialmente, que dirige el foco de atención a zonas concretas del estímulo, y cuyo resultado depende de la tarea que se esté desempeñando.

Sobre esta base (Koch and Ullman, 1985) proponen la formación de una serie de *mapas* mediante mecanismos neuronales en los que se representa el resultado de los procesamientos paralelos de las componentes primarias del estímulo, tales como color, orientación de contornos, disparidad y movimiento. Cabe resaltar que cada uno de estos mapas se calcula de forma independiente, sobre la base experimental de que la saliencia de una determinada característica es independiente de las demás. Este resultado es lo que llaman *mapas de saliencia*, donde básicamente aparecen resaltadas las zonas en las que se producen valores de estas características que no son acordes con los de su entorno, es decir, que de alguna manera resultan extraños en su contexto.

La idea de los mapas de saliencia es recogida en (Itti et al., 1998) y (Itti and Koch, 2001), donde se propone una implementación mediante un programa de computador de este mecanismo atencional, haciendo uso de ella para realizar un análisis, desde el punto de vista de su saliencia, de imágenes. Formalmente, estos mapas se concretan en un valor escalar asociado a cada coordenada 2D del estímulo visual.

Como se ha mencionado, estos trabajos demuestran que la saliencia de un estímulo se relaciona exclusivamente con su *excepcionalidad* en un determinado entorno. La conclusión inmediata es que la saliencia no depende del valor real de la característica, sino de su *contraste* con el entorno; esto conduce a la conclusión de que el mismo valor de una característica puede ser muy saliente en un determinado *contexto* (i.e. combinación de estímulo y tarea), y completamente irrelevante en otro.

En línea con las conclusiones presentadas anteriormente, en (Tversky, 1977) se establece que un estímulo tiene nula saliencia cuando es compartido por toda la población en estudio, pero cuando su valor va siendo más raro (menos probable), su saliencia va aumentando. Este mismo principio vuelve a aparecer en (Itti and Baldi, 2009) para explicar el concepto de *sorpresa*, trabajo en el que además se corrobora la idea de que cuanto menos frecuente es el valor de una característica más información aporta respecto al estímulo que la produce. En (Treue, 2003) esta idea se formaliza estableciendo que mejora su SNR (*Signal to Noise Ratio*).

2.2. El estudio de la similitud

El concepto de similitud es crucial para entender, desde el punto de vista psicológico, muchos de los comportamientos que muestra un ser humano, como son la memoria, la capacidad de establecer una clasificación o la posibilidad de resolver tareas o tomar decisiones. Todos estos aspectos de la conducta humana están determinados por nuestra capacidad para establecer similitudes entre la tarea que nos encontramos resolviendo y otras que hayamos resuelto en el pasado. Se ha afirmado, que la similitud es un concepto presente en casi cualquier aspecto de la psicología (Tversky, 1977). Este es el motivo por el que el estudio de la similitud como comportamiento psicológico ha recibido amplia atención por parte de la comunidad científica.

Existen dos teorías dominantes para explicar la percepción humana de similitud, las conocidas como teoría espacial y teoría de conjuntos de características (Larkey and Markman, 2005). El ejemplo más claro de la primera línea de trabajo es el conocido como MDS (*Multidimensional Scaling*, escalado multidimensional), formalizado en (Shepard, 1962a) y (Shepard, 1962b). El supuesto básico de esta teoría es que la similitud entre dos estímulos visuales es inversamente proporcional a la distancia entre dos puntos en el espacio de características, cada uno definido por un conjunto de descriptores numéricos extraídos de las respectivas imágenes.

Rebatiendo a la aproximación basada en MDS surge la teoría de conjuntos de características, formulada en (Tversky, 1977) y (Tversky and Gati, 1982), y que se fundamenta en la demostración de que la evaluación humana de la similitud no comparte las propiedades fundamentales de cualquier función de distancia, en concreto, las de autosimilitud y minimalidad, simetría y desigualdad triangular. Partiendo de la constatación de este hecho, formula el *modelo de contraste*, originalmente válido sólo para propiedades cualitativas, y que evalúa la similitud mediante la aportación positiva de las propiedades compartidas por ambos estímulos y la aportación negativa de las exclusivas de cada uno de ellos.

En (Ashby and Perrin, 1988) se puede encontrar una excelente revisión de la evolución de ambas aproximaciones.

Dentro del ámbito de los VIRS, se han abordado diferentes aproximaciones al problema de cuantificar la similitud entre dos estímulos representados por un vector de características, siguiendo enfoques de todo tipo, como puedan ser geométricos, probabilísiticos, cualitativos o empíricos. El mas utilizado sigue siendo la aplicación de la aproximación MDS, reflejada en el extensivo uso que se hace de la distancia euclídea para establecer la comparación de similitud entre las características extraídas de dos estímulos. Sobre esta base se han establecido modificaciones orientadas a mejorar su comportamiento, como pueda ser el uso de factores de ponderación obtenidos por aplicación de distintos principios. Esta aproximación al problema se ha reflejado en la bibliografía en multitud de ocasiones, y tanto en el ámbito del puro estudio de la similitud, como queda reflejado en la mencionada recapitulación en (Ashby and Perrin, 1988), como más específicamente en el de la recuperación de imágenes por contenido visual, existiendo una exhaustiva recopilación en (Eidenberger, 2006).

En cuanto a la línea abierta por el trabajo de Tversky, también se han producido aplicaciones en los VIRS, como pueda ser el trabajo reflejado en (Santini and Jain, 1999), donde se persigue igualmente establecer una correcta evaluación de similitud entre dos estímulos sirviéndose de una aproximación *fuzzy* al problema. En el trabajo que ahora se presenta la base para el establecimiento de la medida de similitud está en la modelización y aplicación de un concepto perceptual como es la saliencia.

2.3. *Saliencia y tareas perceptuales*

La idea de que la saliencia es clave para la resolución de tareas perceptuales ha sido utilizada con relativa frecuencia en Psicología. En (Tversky, 1977) ya se refiere que las características salientes de los estímulos implicados es clave para la correcta evaluación de similitud. Este mismo concepto vuelve a aparecer en (Itti and Koch, 2001) donde se demuestra cómo diferentes características de un mismo estímulo contribuyen de distinta forma al resultado final. En (Rao and Ballard, 1999) se confirma este comportamiento mediante el uso de redes neuronales jerárquicas, que se utilizan para generar una codificación predictiva de información presente en el córtex visual utilizando sólo los estímulos con baja probabilidad local, i.e. los más salientes. En esta línea, en (Tsotsos et al., 1995) y (Fairhall et al., 2001) se concluye que el sistema visual humano se centra en las características salientes para resolver tareas.

En cuanto a la cuestión sobre cómo representar la información sobre la tarea en comparación con la del estímulo, en (Itti et al., 1998) se propone que ambas se configuran como *mapas de saliencia*, dejando claro que ambos aspectos del problema (estímulo y tarea) se abstraen de la misma manera a nivel psicológico; así, aparecen los conceptos de saliencia *bottom-up*, asociada y dependiente exclusivamente del estímulo, lo que Tversky llama el *factor intensivo*, y saliencia *top-down*, asociada y dependiente exclusivamente de la tarea que se ha de llevar a cabo, definida por el mismo autor como *factor diagnóstico*; dependiendo de la tarea, la una modula y se impone a la otra.

Queda pendiente establecer la forma en la que la información procedente del estímulo, saliencia bottom-up, y la procedente de la tarea, saliencia top-down, se han de combinar para generar el resultado de la tarea. Aunque no existe una metodología universal para la resolución de esta cuestión (Itti and Koch, 2001), es posible encontrar indicios sobre cómo tratar este problema de forma genérica. Por otro lado, también se ha demostrado (Tversky, 1977) que, hablando ya del estudio de similitud, en un proceso de comparación entre dos estímulos el comportamiento de éstos últimos no es simétrico, en el sentido de que se considera de distinta forma al prototipo que a la instancia, siendo más parecida la segunda al primero que al contrario. En cualquier caso, en los citados trabajos el resultado se expresa con un escalar, que establece el nivel de correlación entre ambas saliencias.

2.4. *Rebatiendo el modelo euclídeo*

Sean I_u e I_v dos imágenes extraídas de un cierto dominio Γ, siendo $\mathbf{X}_u = \{x_{u1}, x_{u2}, \ldots, x_{uN}\}$ y $\mathbf{X}_v = \{x_{v1}, x_{v2}, \ldots, x_{vN}\}$ los vectores de características extraídos, respectivamente, de cada una de ellas, con $\mathbf{X}_{\{u,v\}} \in \Omega \subset \mathbb{R}^N$, donde Ω representa el subespacio en el que los vectores de características extraídos de las imágenes en Γ pueden tomar valores. La forma más extendida en los VIRS para establecer una comparación entre ambos vectores que lleve a establecer su similitud (Santini and Jain, 1999) (Ashby and Perrin, 1988), es mediante el uso de la distancia de Minkowsky aplicada a los puntos resultantes de considerar cada uno de estos vectores de características como unas coordenadas en el espacio de características N-dimensional, práctica que se fundamenta en los estudios publicados en (Shepard, 1962a) y

(Shepard, 1962b):

$$D_m(\mathbf{X}_u, \mathbf{X}_v) = \left(\sum_{i=1}^{N} (x_{ui} - x_{vi})^m \right)^{\frac{1}{m}} \quad (1)$$

Debido a esta amplia implantación, como ya se ha mencionado, esta es la referencia sobre la que se discute el modelo de evaluación de similitud presentado en este trabajo.

A partir de (1) es evidente que todas las componentes del vector de características son ponderadas igualmente en el proceso de comparación, lo que no hace sino reflejar la hipótesis subyacente de que todas ellas son iguamente relevantes para establecer la similitud entre ambos estímulos (suponiendo, lógicamente, que todas ellas han sido normalizadas para compatibilizar adecuadamente sus rangos). Sin embargo, esta hipótesis entra en conflicto con los resultados presentados en las secciones anteriores. Por tanto, este equilibrio en la ponderación de las características usadas en la medida de similitud debe abandonarse.

Es fácil encontrar ejemplos que apoyan estas conclusiones. En Tversky (1977) pueden encontrarse algunos casos sencillos aplicados a imágenes, y validados por la experiencia con observadores voluntarios. Supóngase el caso de clasificar especies animales atendiendo a algunas de sus medidas anatómicas. Se presentan distintos casos donde una única característica puede resolver la tarea por si misma, prácticamente descartando la información aportada por el resto del vector de características:

- Pantera vs. leopardo: para un sujeto poco avezado en zoología, las medidas anatómicas de ambas especies pueden no resultar concluyentes (tamaño y peso similares, piezas dentales casi iguales...). Sin embargo, haciendo uso exclusivamente del aspecto de su piel, muy distinta en ambos casos, determinar que pertenecen a especies distintas resulta trivial.
- leopardo adulto vs. leopardo joven: en este otro problema, la situación es distinta, al ser prácticamente todas las medidas anatómicas diferentes por el distinto grado de madurez de ambos ejemplares. No obstante, la igualdad en este caso del dibujo de la piel permite alcanzar con facilidad la conclusión correcta de que pertenecen a la misma especie.
- Modificando ahora la tarea, se pide establecer una ordenación atendiendo al parecido entre un leopardo adulto (utilizado como referente de la comparación) y los otros dos mencionados ejemplares, pantera y leopardo joven. Aunque casi todas las medidas anatómicas apuntan a que leopardo adulto y pantera son animales prácticamente iguales, la evaluación del patrón de piel probablemente conducirá a revocar esta conclusión, asignando más parecido entre el ejemplar joven y el adulto en razón de pertenecer a la misma especie.

La conclusión a la que nos lleva esta serie de ejemplos es que la piel del ejemplar es una característica definiva para completar la tarea, prácticamente descartando la información proporcionada por el resto de características. Dicho de otra forma, en este contexto, es la característica más relevante. No obstante, en otra tarea, como puede ser determinar si dos ejemplares son del mismo género, el patrón de la piel pasa a ser trivial, debiendo fijar la atención en otros parámetros como peso, dimensiones, etc. Resumiendo los casos presentados, una característica saliente permite resolver correctamente un problema ...

- ...de clasificación donde las demás características conducen a un falso positivo.
- ...de clasificación donde las demás características conducen a un falso negativo.
- ...de ordenación donde las demás características conducen a un error

y sin embargo, no permite resolver problemas cuando es compartida por ambos estímulos, lo que demuestra que su relevancia depende del contexto en el que se utiliza.

Siguiendo este razonamiento, se presenta a continuación una metodología para incorporar la información de saliencia de una determinada característica en los procesos de evaluación de similitud entre dos estímulos.

3. Evaluación de similitud

En esta Sección se presenta el modelo de evaluación de similitud, que siguiendo una estrategia novedosa, establece una ponderación para cada característica utilizada en un proceso de comparación. A diferencia de los trabajos anteriores en esta dirección, la metodología utilizada se basa exclusivamente en la incorporación a los cálculos de una cuantificación de un mecanismo clave en la explicación de nuestro comportamiento perceptual en el desarrollo de esta tarea, como es la propia saliencia, y no simplemente en la incorporación de un formalismo matemático, ya sea fuzzy (Santini and Jain, 1999), geométrico o estadístico (Eidenberger, 2006).

3.1. El factor de saliencia, SF(**X**)

En esta sección se presenta el mecanismo propuesto para la cuantificación de la saliencia de una característica dada dentro de un vector de características, que se corresponde con el factor intensivo (Tversky, 1977) o la saliencia bottom-up (Itti et al., 1998). Para ello se introduce la siguiente nomenclatura; sean:

- $I \in \Gamma$, una imagen perteneciente al dominio de extracción definido
- $\mathbf{X} = \{x_1, x_2, \ldots, x_N\} \in \Omega \subset \mathbb{R}^N$, el vector de características asociado a dicha imagen, definido en un subespacio Ω dentro del espacio vectorial de dimensión N.
- $\mathbf{P}(\mathbf{X}) = \{P_1(x_1), P_2(x_2), \ldots, P_N(x_N)\} = \{p_1, p_2, \ldots, p_N\} \in [0,1]^N$ el vector que contiene las probabilidades de aparición de los valores de cada característica. La notación $P_i(x_i)$ hace referencia a que la probabilidad se ha de calcular separadamente para cada característica.

Dadas estas definiciones, se supone como hipótesis de partida que el conjunto de caracterísitcas $x_i, i = 1 \ldots, N$ contiene la suficiente información como para resolver la tarea, es decir, queda fuera del ámbito de este trabajo la tarea de selección de características.

Dados estos elementos se propone la existencia del *Factor de Saliencia* $SF(x_i) = sf_i$ asociado a una característica x_i, y cuyo comportamiento viene definido por el siguiente conjunto de proposiciones:

Proposición 3.1. *Sea* $\mathbf{X} = \{x_1, x_2, \ldots, x_N\} \in \Omega \subset \mathbb{R}^N$ *el vector de características extraído de la imagen* $I \in \Gamma$. *Entonces, existe un escalar, llamado factor de saliencia asociado a la característica* $x_i, i = 1, \ldots, N$, $SF(x_i) = sf_i$ *que cuantifica la saliencia de dicha característica, tal que:*

$$SF : \mathbb{R} \longrightarrow \mathbb{R} \tag{2}$$

A la vista de la proposición 3.1, el factor de saliencia se materializa en un valor escalar que refleja la relevancia de una característica en cuanto a la conformación del estímulo, tal como sucede en los trabajos citados en la sección 2.1. La hipótesis que se aporta originalmente en el presente trabajo es la de sustituir el concepto de contexto: originalmente se asocia a un entorno local en el estímulo visual, es decir, a una determinada región de la imagen; en este trabajo se entiende contexto como el conjunto de valores que alcanza una característica. Por lo tanto si una zona saliente de una imagen es aquélla que se manifiesta distinta a la región que la contiene, un valor saliente será aquél que se manifieste como poco común para una característica.

Como se explica también en la sección 2.1, el valor de la saliencia se relaciona por tanto de forma directa con la baja probabilidad de que una determinada característica tome un determinado valor en un contexto. Siguiendo este razonamiento, se puede enunciar la siguiente proposición:

Proposición 3.2. *El factor de saliencia asociado al valor de una característica* $x_i, i = 1, \ldots, N, SF(x_i) = sf_i$, *depende exclusivamente de la probabilidad de que dicha característica tome su valor actual,* $P_i(x_i) = p_i$, *y por lo tanto se puede ajustar su definición a:*

$$SF : [0, 1] \longrightarrow \mathbb{R} \tag{3}$$

A la vista de esta propiedad está claro que debería cambiarse la notación a $SF(p_i)$ en lugar de $SF(x_i)$. Sin embargo, se mantendrá esta última para expresar la asociación entre el factor de saliencia y la característica que lo origina.

Las siguientes proposiciones introducen las ideas de cuantificación de saliencia mencionadas en el estado del arte, en referencia a su anulación para valores de características compartidos por la mayoría de la población, y su máxima influencia cuando son exclusivos, presentes en muy pocos individuos. En ellas se intoduce en el modelo la no negatividad y el no crecimiento para el factor de saliencia.

Proposición 3.3. *El factor de saliencia asociado al valor de la característica* $x_i, i = 1, \ldots, N, SF(x_i) = sf_i$, *es un número real y no negativo, por lo que la función* $SF(x_i) = sf_i$ *debe ser semidefinida positiva:*

$$SF : [0, 1] \longrightarrow \mathbb{R}^+ \tag{4}$$

Proposición 3.4. *Suponiendo derivable al factor de saliencia asociado al valor de la característica* $x_i, i = 1, \ldots, N, SF(x_i) = sf_i$, *entonces:*

$$\frac{d\ sf_i}{d\ p_i} \leq 0,\ \forall p_i \in [0, 1] \tag{5}$$

En caso contrario, debe cumplir la condición de ser una función no creciente en dicho intervalo.

A partir de estas proposiciones, y siguiendo las hipótesis que formalizan, la elección de la forma del factor de saliencia es dependiente de la aplicación, y debe reflejar la importancia asignada a la excepcionalidad en el proceso de comparación.

3.2. Comparación de dos imágenes: el factor de coincidencia $CF(x_{ui}, x_{vi})$

El razonamiento se centra ahora en la definición de la tarea en la que dicho estímulo se verá inmerso, es decir, un proceso de evaluación de la similitud o diferencia entre las características extraídas de dos estímulos. Para ello, se aplica el principio de que tanto el estímulo como la tarea se codifican en forma de valores de saliencia, por lo que se propone un mecanismo para su combinación, uno procedente del estímulo que se pretende comparar (factor intensivo) y otro procedente del referente, que se utiliza como codificación de la tarea (factor diagnóstico).

Se propone ahora una metodología original para establecer la combinación de la información procedente de ambos estímulos en un proceso de evaluación de similitud. En lo sucesivo se utiliza la siguiente nomenclatura:

- $\{I_u, I_v\} \in \Gamma$ un par de imágenes que se someterán al proceso de evaluación de similitud, ambas extraídas del mismo dominio
- $\mathbf{X}_u = \{x_{u1}, \ldots, x_{uN}\}$ y $\mathbf{X}_v = \{x_{v1}, \ldots, x_{vN}\}$, $\{\mathbf{X}_u, \mathbf{X}_v\} \in \Omega \subset \mathbb{R}^N$, los vectores de características extraídos de cada una de ellas respectivamente
- $SF(\mathbf{X}_u) = \{sf_{u1}, \ldots, sf_{uN}\}$ y $SF(\mathbf{X}_v) = \{sf_{v1}, \ldots, sf_{vN}\}$, $\{SF(\mathbf{X}_u), SF(\mathbf{X}_v)\} \in \mathbb{R}^{N+}$, los vectores que contienen los factores de saliencia asociados a las valores de las características de cada uno.

Se propone ahora la existencia de un *Factor de Coincidencia*, $CF(x_{ui}, x_{vi}) = cf_i^{uv}$, generado por un par de valores de una característica x_i medida en dos imágenes sometidas a un proceso de evaluación de similitud:

Proposición 3.5. *Sean* $\{I_u, I_v\} \in \Gamma$ *dos imágenes sometidas a un proceso de evaluación de similitud, y* $\{\mathbf{X}_u, \mathbf{X}_v\} \in \Omega \subset \mathbb{R}^N$ *los*

vectores de características extraídos a partir de ellas, respectivamente. Entonces existe un escalar llamado factor de coincidencia para la característica i-ésima, que pondera su relevancia combinando los factores de saliencia del par de valores de dicha característica, $\{x_{ui}, x_{vi}\}$, y que se define como:

$$CF : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R} \tag{6}$$

El factor de coincidencia debe reflejar la forma en la que se entiende que la información asociada a cada par de características se debe combinar para resolver la tarea. Dado que en este trabajo la hipótesis fundamental es que dicha información se encuentra recogida en los factores de saliencia, el factor de coincidencia debe basarse en ellos. Por lo tanto, se formula la siguiente proposición.

Proposición 3.6. *El factor de coincidencia para un par de valores (x_{ui}, x_{vi}), $CF(x_{ui}, x_{vi}) = cf_i^{uv}$, depende de los factores de saliencia de ambos, sf_{ui} y sf_{vi}, respectivamente. Por lo tanto, se puede modificar su definición como:*

$$CF : \mathbb{R}^+ \times \mathbb{R}^+ \longrightarrow \mathbb{R} \tag{7}$$

A la vista de esta proposición, parece claro que la notación debería cambiar de $CF(x_{ui}, x_{vi})$ a $CF(sf_{ui}, sf_{vi})$; no obstante, con objeto de mantener la asociación entre la definición del factor de coincidencia y los valores de las características que lo originan se mantiene sin modificaciones.

Al igual que sucede con el factor de saliencia, se pueden fijar los criterios directores que deben guiar la decisión de cómo concretar el factor de coincidencia entre dos estímulos. En primer lugar, es obvio que cuando el valor de una característica tiene una saliencia baja para cualquiera de las dos imágenes en comparación, el peso relativo de esta característica en el proceso debe ser bajo. En el límite, si ambos factores de saliencia fueran nulos, su factor de coincidencia debería ser también nulo. En el otro extremo, cuando ambas características tengan factores de saliencia elevados su factor de coincidencia debe ser asimismo elevado, puesto que sus valores representan a un término muy distintivo para ambos estímulos.

Este razonamiento se concreta en las dos siguientes proposiciones:

Proposición 3.7. *El valor del factor de coincidencia del par $\{x_{ui}, x_{vi}\}$, $CF(x_{ui}, x_{vi}) = cf_i^{uv}$, es un número real no negativo, así que la función que lo define debe ser semidefinida positiva. De esta forma, se puede modificar su definición:*

$$CF : \mathbb{R}^+ \times \mathbb{R}^+ \longrightarrow \mathbb{R}^+ \tag{8}$$

Proposición 3.8. *Si la función que define el factor de coincidencia del par $\{x_{ui}, x_{vi}\}$, $CF(x_{ui}, x_{vi}) = cf_i^{uv}$ es derivable se debe cumplir que:*

$$\frac{\partial cf_i^{uv}}{\partial sf_{ki}} \geq 0, \qquad k \in \{u, v\} \tag{9}$$

En caso contrario, debe ser una función no decreciente con respecto a sus dos parámetros.

Enlazando las proposiciones 3.4 y 3.8 se llega a:

$$\frac{\partial cf_i^{uv}}{\partial p_{ki}} = \frac{\partial cf_i^{uv}}{\partial sf_{ki}} \frac{\partial sf_{ki}}{\partial p_{ki}} \leq 0, \quad k \in \{u, v\} \tag{10}$$

que explica la dependencia entre el factor de coincidencia y la probabilidad de alcanzar un determinado valor para una característica. Esta ecuación muestra que a medida que la probabilidad de los valores en comparación aumenta, éstos tienen una carga descriptiva inferior a la de aquéllos poco frecuentes dentro del universo del experimento.

3.3. Medida de similitud basada en la saliencia

En la sección anterior se ha introducido el concepto de factor de coincidencia como elemento que cuantifica la relevancia de cada uno de los términos (valores de características) que se utilizan en un proceso de comparación. Este factor de coincidencia se basa en los respectivos factores de saliencia, que representan la información intensiva y diagnóstica para la realización de la tarea. A partir de su valor, en este trabajo se propone el establecimiento de un mecanismo de ponderación de cada comparación individual de características, que en el caso de utilizar como base la distancia euclídea se concreta en la siguiente expresión:

$$MD_2(\mathbf{X}_u, \mathbf{X}_v) = \left(\sum_{i=1}^{N} cf_i^{uv} |x_{ui} - x_{vi}|^2 \right)^{\frac{1}{2}} \tag{11}$$

4. Resultados

La medida de similitud presentada en este trabajo se ha probado en una tarea real de recuperación de imágenes por contenido. En ella, se parte de una población de más de 31.000 imágenes[1]. Todas ellas son binarias, y en la población de la base de datos no se han aplicado restricciones de tamaño ni resolución. Su aspecto es exactamente igual al de la imagen que se muestra en la figura 1.

Se ha decidido utilizar esta base de datos ad-hoc en lugar de alguna de las existentes para benchmarking fundamentalmente por cuestiones de adecuación al problema: objetos sencillos de formas claramente definidas y fácil caracterización y en número suficiente como para incluir parecidos, distractores y una escala similar a la de un problema real de recuperación. De todas las consultadas, la más adecuada según estos requisitos fue la base SIID de la Universidad de Brown; sin embargo, su reducido tamaño hizo descartar esta opción en favor de nuestra propuesta. Una recopilación de bases de datos de imágenes para benchmarking en VIRS se puede encontrar en (Fisher, 2011).

[1] disponibles mediante solicitud al autor

Figura 1: Aspecto de las imágenes utilizadas en el experimento

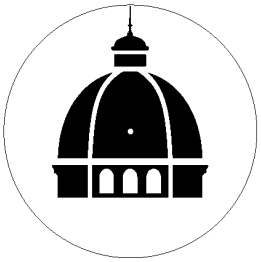


Figura 2: Resultado del proceso de normalización de las imágenes: el centroide del campo negro se representa con un pequeño círculo blanco, mientras que el círculo unidad que rodea la figura aparece en línea continua, siendo tangente al punto de la figura más alejado del centroide. Nótese que los píxeles exteriores a este círculo no serán utilizados posteriormente en la etapa de caracterización.

4.1. *Procesamiento de las imágenes*

Con objeto de hacer compatible la representación de las imágenes, se ha sometido a todas ellas a un proceso de normalización, consistente en las siguientes operaciones:

- el centroide de la figura que aparece en la imagen (campo negro) se ha hecho coincidir con el centro de la imagen
- la figura se ha escalado para que el punto más alejado del centroide sea tangente a la máxima circunferencia interior a la imagen
- se ha recortado la imagen para que dicha circunferencia sea tangente interior a sus límites (i.e. se ha generado la imagen cuadrada más pequeña que circunscribe a la circunferencia)
- se ha escalado la imagen así generada a un tamaño fijo de 128×128 píxeles.

Se eligió este proceso de normalización, basado en la definición de un círculo *unidad*, dado que la caracterización de las imágenes, como se explicará a continuación, hace uso de descriptores cuyo cálculo se realiza en coordenadas polares. Obviamente, esto hace que los píxeles exteriores a este círculo unitario no se utilicen para el cómputo de estas características. En cuanto al tamaño final de las imágenes se ha elegido de forma experimental para compatibilizar una resolución suficiente en la descripción de las figuras con una moderada carga computacional en el procesamiento de cada una. El resultado de este procesamiento puede observarse en la figura 2.

4.2. *Extracción de características*

Las imágenes completas han sido descritas mediante momentos de Zernike; este descriptor ha sido usado frecuentemente para la caracterización de imágenes como las utilizadas en este experimento (Kim and Kim, 1998)(Kim et al., 2000), dada su invarianza ante rotación, que se une a la invarianza ante traslación y escala conseguidas mediante el proceso de normalización de las imágenes explicado anteriormente. En (Teh and Chin, 1988) se demuestra su superioridad frente a otras descripciones basadas en momentos mediante un estudio comparativo, y continúan utilizándose como referencia a la hora de evaluar nuevos métodos de descripción de regiones, como en (Chen and Xie, 2011). Su definición se basa en los polinomios de Zernike, de la forma:

$$R_{nm}(\rho) = \sum_{s=0}^{\frac{n-|m|}{2}} (-1)^s \frac{(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!}\rho^{n-2s} \qquad (12)$$

$$R(\rho) = \{R_{nm}(\rho) \mid n = 0, 1, \ldots, \infty, \quad |m| \leq n, \quad n - |m| \text{ par}\}$$

Basándose en la familia de polinomios descrita en (12) se definen los momentos bidimensionales de Zernike de una imagen $I(\rho, \theta)$ (i.e. representada en coordenadas polares, frente a la habitual representación en cartesianas $I(i, j)$) como:

$$A = \frac{n+1}{\pi} \sum_{\rho} \sum_{\theta} [V(\rho, \theta)]^* I(\rho, \theta), \quad \rho \leq 1 \qquad (13)$$

donde $[.]^*$ representa la transpuesta conjugada del vector, y:

$$V(\rho, \theta) = R(\rho) e^{-im\theta} \qquad (14)$$

Es interesante mencionar algunos aspectos sobre el uso de momentos de Zernike:

- Cada momento extraído de la imagen es un número complejo; en nuestra aplicación sólo se utiliza el módulo, de forma que el vector de características empleado es:

$$\mathbf{X} = |A| \qquad (15)$$

- Los polinomios de Zernike forman una base ortonormal, por lo que es posible realizar una reconstrucción de la imagen original a partir del conjunto de momentos extraídos. Esta propiedad se ha utilizado para tomar la decisión acerca de cuál debe ser el orden en el que se trunque la serie de momentos para obtener una reconstrucción suficientemente detallada de la imagen procesada. En este caso, la decisión ha sido alcanzar hasta momentos de orden $n = 17$, lo que da un vector de características de 90 componentes.

- Por definición (13), los momentos de Zernike se calculan en el interior del círculo unidad, por lo que se ha escalado la medida del radio real en la imagen, $r \in [0, 64]$, para cumplir con esta condición haciendo que $\rho \in [0, 1]$.

Una vez extraídos todos los momentos, el valor de cada uno de ellos ha sido reescalado para equilibrar su contribución a la medida de similitud en el proceso de comparación. Así, el vector de características que se utiliza se define como:

$$\tilde{\mathbf{X}} = \{\tilde{x}_i \mid \tilde{x}_i \in [0, 1], \quad i = 1, \ldots, N\} \tag{16}$$

siendo

$$\tilde{x}_i = \frac{x_i - \text{mín}(x_{ki})}{\text{máx}(x_{ki}) - \text{mín}(x_{ki})}, \quad \forall\{k : I_k \in \Gamma\} \tag{17}$$

4.3. *Estimación de probabilidades*

Una vez que se dispone de los vectores de características, se realiza la estimación de la probabilidad de cada una de sus componentes, que como se ha explicado servirá después para realizar el cálculo del factor de saliencia de cada una de ellas. Para realizar esta estimación se ha seguido el siguiente procedimiento:

- para cada característica $\tilde{x}_i$ se ha recopilado su valor en cada una de las imágenes utilizadas en el experimento, formando el conjunto:

$$\{\tilde{x}_i\} = \{\tilde{x}_{ki} \mid \forall I_k \in \Gamma\} \tag{18}$$

- a partir de cada conjunto $\{\tilde{x}_i\}$ así formado, se ha generado su histograma, H_i, fijando experimentalmente el ancho de cada uno de sus *bins* en $\Delta\tilde{x}_i = 0{,}1$, de tal forma que se generan diez niveles; este número de niveles representa un compromiso entre una cantidad suficientemente alta para diferenciar entre rangos de valores del parámetro y lo suficientemente bajo para que haya diferencias significativas en el número de individuos asociados a cada tramo.

- la estimación de la probabilidad $P(\tilde{x}_i) = p_i$ se toma como el área relativa del *bin* en el que el valor $\tilde{x}_i$ estaría alojado en el histrograma H_i correspondiente a cada conjunto $\{\tilde{x}_i\}$.

Para verificar si con las probabilidades calculadas el mecanismo de aplicación de la saliencia a la ponderación de la medida de similitud resultaría efectivo, se analizó la distribución de probabilidades para cada una de las características, por evaluación de la forma de los distintos histogramas generados, encontrándose que la tendencia es a que adopten la forma de distribuciones Gamma. Este hecho confirma la observación recogida en (Kim and Kim, 1998), y marca diferencias significativas de probabilidad en distintos rangos de valores para cada característica, lo que a su vez generará diferencias apreciables en los factores de ponderación aplicados. Nótese que distrubiciones uniformes hubieran resultado en evaluaciones de similitud virtualmente iguales a las obtenidas con la distancia euclídea.

4.4. *Definición de la medida de similitud*

Para poder establecer la capacidad de la medida de similitud en relación con la distancia euclídea, es necesario fijar la forma que toman tanto el factor de saliencia como el factor de coincidencia. En el experimento cuyos resultados se presentan, se ha optado por utilizar la misma definición para los factores de saliencia correspondientes a todas las características. No obstante, en aplicaciones en las que existan diferencias cualitativas entre las características utilizadas, sería posible diferenciar cada uno de estos factores para cada una de ellas. En concreto, en este caso se ha definido el factor de saliencia:

$$SF(x_i) = sf_i = 1 - P(\tilde{x}_i) = 1 - p_i, \qquad \forall i = 1, \ldots, N \tag{19}$$

definición que obviamente cumple con las proposiciones referidas.

En cuanto al factor de coincidencia utilizado, se ha definido como:

$$CF(\tilde{x}_{ui}, \tilde{x}_{vi}) = cf_i^{uv} = sf_{ui}\ sf_{vi} = (1-p_{ui})(1-p_{vi}), \quad \forall i = 1, \ldots, N \tag{20}$$

Nótese que estas definiciones describen un comportamiento para el factor de ponderación caracterizado por:

- La medida de similitud que se aplica con el factor de coincidencia definido en (20) tiene como cota superior el valor de la distancia euclídea medida sobre los mismos estímulos en comparación:

$$\left.\begin{array}{l} 0 \le SF(\tilde{x}_{ui}) \le 1 \\ 0 \le SF(\tilde{x}_{vi}) \le 1 \end{array}\right\} \rightarrow 0 \le CF(\tilde{x}_{ui}, \tilde{x}_{vi}) \le 1 \tag{21}$$

$$\Rightarrow MD_2(\tilde{\mathbf{X}}_u, \tilde{\mathbf{X}}_v) \le D_2(\tilde{\mathbf{X}}_u, \tilde{\mathbf{X}}_v) \tag{22}$$

  donde $D_2(.)$ representa la distancia euclídea.

  Gracias a esta propiedad adicional, en el experimento que se va a describir ha sido posible cuantificar el efecto de la aplicación de la corrección a través de los factores de ponderación incluidos en la medida de similitud, por simple observación de la cuantía en que dicha medida decrece con respecto al valor de distancia calculado.

- De otra parte, se ha construido el factor de ponderación considerando por igual los factores de saliencia procedentes de ambos estímulos, es decir, que se ha tenido en cuenta de igual forma la saliencia bottom-up (dependiente del estímulo) y la saliencia top-down (dependiente de la tarea, en este caso el referente de la comparación). Esta elección se debe a la concurrencia de dos circunstancias: por un lado, como ya se ha mencionado, no existen directrices en la bibliografía sobre cómo construir un mapa de saliencia top-down; por otra parte, lo normal en un proceso de comparación de dos estímulos para evaluar su parecido, es considerar ambos por igual, sin establecer diferencias sobre cuál representa el query y cuál está incluido en la base de datos.

$$CF(x_{ui}, x_{vi}) = CF(x_{vi}, x_{ui}) \tag{23}$$

Esta es una decisión de diseño del experimento que no tiene por qué producirse de igual forma en otros contextos o para otras tareas. En circunstancias en las que exista una clara asimetría en el comportamiento de los estímulos que se sometenten a comparación, se diseñaría un factor de coincidencia que materializaría la dominancia de uno de ellos.

4.5. *Diseño del experimento*

Para la prueba de la medida de similitud, se seleccionaron al azar 100 imágenes, que fueron utilizadas como queries a la base de datos completa. Dado que el objetivo del expermento no es comprobar la capacidad de recuperar variaciones de la imagen utilizada para el query (problema de *matching*), no se realizó ninguna modificación sobre ninguna de ellas, dando por seguro que la identidad se recupera en primer lugar en todos los casos.

En lo sucesivo, se va a utilizar la siguiente notación para referirse a un conjunto ordenado de imágenes recuperadas como consecuencia de un query. Sea la imagen de consulta $I_q \in \Gamma$, sea ω el orden de los momentos de Zernike utilizados en la comparación, y M el cardinal del conjunto de imágenes recuperadas; entonces, el conjunto ordenado de imágenes recuperadas se denota por $I_Q(\omega, \Phi(.,.), M)$, donde $\Phi(.,.)$ es la función utilizada para la evaluación de la similitud, y que en el caso del experimento que se describe puede ser la distancia euclídea, en cuyo caso $\Phi(.,.) = D_2(.,.)$, o bien la medida de similitud propuesta, $\Phi(.,.) = MD_2(.,.)$. Al ser ordenados, cualquier elemento de cada uno de estos conjuntos posee un subíndice que indica su posición dentro del conjunto: a menor subíndice, mayor similitud con el referente (i.e. imagen de query).

Para cada una de estas 100 queries se recuperaron dos conjuntos de imágenes de 25 elementos cada uno, el primero utilizando la distancia euclídea en la comparación, y el segundo la medida de similitud; siguiendo la notación propuesta, los conjuntos recuperados son, pues:

$$\left.\begin{array}{l} \{I_Q(17, D_2(.,.), 25\} \\ \{I_Q(17, MD_2(.,.), 25\} \end{array}\right\} Q = 1, \ldots, 100 \qquad (24)$$

Una vez recuperados estos conjuntos de imágenes haciendo uso de las dos medidas para la comparación, se contrastaron ambos resultados con la percepción humana de similitud, mediante la participación de voluntarios que se sometieron a la tarea de clasificar como 'Parecida' o 'No parecida' cada una de las imágenes recuperadas para cada consulta en ambos casos. El procedimiento seguido para esta validación se atuvo a los siguientes principios:

- Cada voluntario revisó diez conjuntos de 25 imágenes, correspondientes a 10 queries distintos (i.e. revisión de 250 imágenes)
- Ningún voluntario revisó los dos resultados generados para el mismo query
- Los diez conjuntos se dividieron en cinco correspondientes a distancia euclídea y cinco a medida de similitud, con objeto de no sesgar el resultado a favor de ningíun procedimiento
- Ningún voluntario conocía el objetivo final del experimento
- La evaluación fue siempre cualitativa, como se ha descrito. No se solicitó ninguna evaluación cuantitativa.
- Los conjuntos recuperados se presentaron desordenados, para evitar sesgo a favor de los primeros
- Cada conjunto recuperado fue evaluado una única vez

Durante el experimento no se ha dispuesto de un conjunto etiquetado como *ground truth* para ninguna de las consultas, debido a la escala del problema: 100 consultas a una base de datos de 31000 individuos supone abordar el etiquetado previo de más de tres millones de imágenes. Si se realizara esta tarea consumiendo un segundo por cada imagen se tardarían casi 900 horas de trabajo para disponer del conjunto etiquetado completamente.

Como resultado del proceso de validación, se dispone de una etiqueta para cada imagen recuperada que la señala como parecida o no a la correspondiente imagen de consulta.

4.6. *Descripción de los resultados*

Para poder establecer la comparación entre el conjunto de resultados obtenido con la distancia euclídea y los conseguidos con la medida de similitud, se ha recurrido a calcular la precisión media alcanzada con cada método. La precisión en un conjunto de recuperación se calcula como:

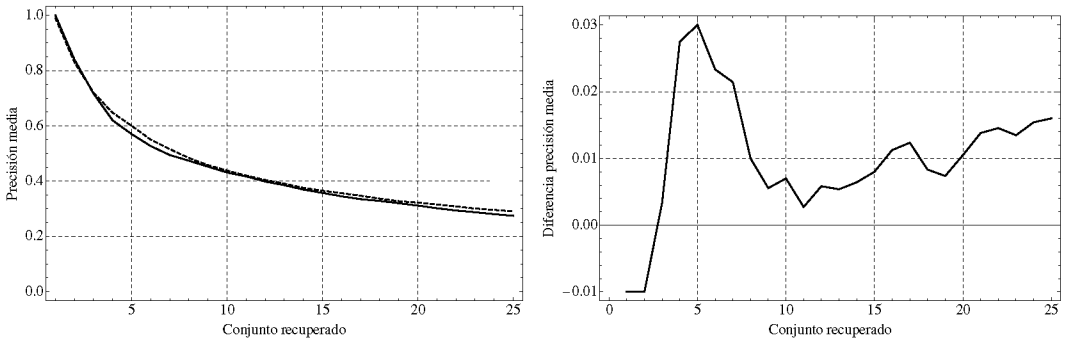
$$\text{precisión} = \frac{\text{imágenes correctamente recuperadas}}{\text{tamaño del conjunto recuperado}} \qquad (25)$$

Nótese que al no poder contar con un conjunto de etiquetas a priori generadas de acuerdo con la relevancia o no de una imagen para cada consulta, no es posible establecer el *recall*, la otra medida clásica utilizada en la evaluación de sistemas de recuperación de información, que se define como:

$$\textit{recall} = \frac{\text{imágenes correctamente recuperadas}}{\text{imágenes similares en la base de datos}} \qquad (26)$$

A partir de la definición (25) el procedimiento para evaluar el rendimiento de cada medida de comparación empleada ha consisitido en calcular para cada consulta y cada procedimiento de comparación la precisión alcanzada para cada tamaño del conjunto de recuperación entre 1 y 25. Una vez que se ha obtenido este valor, se ha promediado cada uno de ellos empleando las 100 consultas realizadas. Al final de este proceso se dispone, para cada procedimiento de comparación, de un conjunto ordenado de 25 valores de precisión media, correspondientes a los promedios de precisión alcanzados para cada tamaño del conjunto recuperado.

Los resultados obtenidos para ambas series de valores se muestran en la figura 3(a); en ella se puede apreciar el típico



(a) Comportamiento neto de la precisión para distancia euclídea (continua) y medida de similitud (discontinua)

(b) Diferencia entre la precisión de la medida de similitud y de la distancia euclidea

Figura 3: Resultados de la validación del expermiento.

comportamiento de una curva de precisón, decreciendo a medida que el conjunto de recuperación aumenta, dando lugar a la inclusión de más falsos positivos.

En la figura 3(a) puede aprecierse como, aun cuando ambas curvas siguen claramente la misma tendencia de precisiones decrecientes, la gráfica en línea discontinua correspondiente a la medida de similitud se sitúa por encima de la de la distancia euclídea para prácticamente todos los tamaños del conjunto de recuperación. Para apreciar esta tendencia con más detalle, se ha incluido la figura 3(b), en la que se han representado las diferencias entre ambas series de valores, mostrándose casi para cada conjunto de recuperación favorable a la medida de similitud propuesta.

En la figura 3(b) se pueden apreciar tres hitos que merece la pena comentar en detalle:

- En primer lugar, y para un tamaño del conjunto de recuperación de dos elementos, se puede apreciar un pico negativo, indicando que el correspondiente valor de precisón media es superior para la distancia euclídea. Esto indicaría un mejor comportamiento para conjuntos de recuperación muy reducidos por parte de ésta. Para comprobar si realmente esta tendencia es estadísticamente significativa, se ha realizado un test del signo de Fischer para realizar el contraste de hipótesis (Urdan, 2005). Para realizar el test, se han tomado los valores individuales de precisión alcanzados para un conjunto de recuperación de dos elementos, para cada una de las 100 consultas realizadas, tanto en el caso de la distancia euclídea como en el caso de la medida de similitud. Esto arroja pares de valores, correspondientes a las precisiones alcanzadas con ambos procedimientos para cada query. El test de signo indica si estos pares provienen de la misma función de distribución de probabilidad, o si por el contrario proceden de dos distintas; esto último indicaría que estadísticamente hablando, uno de los dos métodos, en este caso la distancia euclídea, tiene un mejor comportamiento que el otro. La hipótesis nula es que ambas distribuciones son la misma, y que por lo tanto ningún método es mejor que el otro. Realizado este test, el estadístico obtenido es $\rho = 0{,}6875$, que confirma la hipótesis nula e indica una muy elevada probabilidad de que ambos conjuntos de datos, procedan de la misma función de distribución de probabilidad. De hecho, el rechazo de la hipótesis nula se realiza cuando el estadístico obtenido es $\rho = 0{,}05$ o menor, por lo que claramente queda probado que esta tendencia es un suceso puramente aleatorio.

- A partir de este primer valor negativo, el gráfico evoluciona rápidamente hasta alcanzar un pico positivo para un tamaño de cinco elementos en el conjunto de recuperación, en el que la medida de similitud propuesta se comporta claramente mejor que la distancia euclídea. Al igual que en el caso anterior, se ha validado esta hipótesis con el test del signo, arrojando un valor para el estadístico de $\rho = 0{,}043$, que queda por debajo del umbral de aceptación de la hipótesis nula. Se puede concluir, por lo tanto, que este pico es significativo desde el punto de vista estadístico, y demuestra un mejor comportamiento de la medida de similitud. En concreto esta mejora se puede cuantificar a partir de los valores precisos alcanzados para cada serie:
  - la precisión media para la distancia euclídea es de $\Pi_D = 0{,}57$
  - la precisión media para la medida de similitud es de $\Pi_{MD} = 0{,}6$
  - la mejora de precisión en términos absolutos es por tanto de un $\Delta\Pi = 3\,\%$
  - la mejora relativa de precisión es:

$$\Delta_{rel}\Pi = \frac{\Pi_{MD} - \Pi_D}{\text{máx}(\Pi_{MD}, \Pi_D)} = \frac{0{,}6 - 0{,}57}{0{,}6} = 5\,\% \tag{27}$$

- Tras este máximo local, la serie de diferencias se mantiene siempre en valores positivos, lo que significa que hay una tendencia mantenida a que la medida de similitud tenga mejor comportamiento que la distancia euclídea. Esta tendencia culmina con un nuevo máximo local para un tamaño del conjunto recuperado de 25 elementos. Una

vez más, se valida la relevancia estadística de esta diferencia con el test del signo, obteniéndose un valor para el estadístico de $\rho = 0{,}002$, muy por debajo del umbral de aceptación, lo que indica que claramente la medida de similitud tiene un comportamiento estadísticamente mejor que la distancia. Al igual que en el caso anterior, se puede cuantificar esta mejora con los siguientes datos:

- la precisión media para la distancia euclídea es de $\Pi_D = 0{,}2744$
- la precisión media para la medida de similitud es de $\Pi_{MD} = 0{,}2904$
- la mejora de precisión en términos absolutos es por tanto de un $\Delta\Pi = 1{,}6\,\%$
- la mejora relativa de precisión es:

$$\Delta_{rel}\Pi = \frac{\Pi_{MD} - \Pi_D}{\text{máx}(\Pi_{MD}, \Pi_D)} = \frac{0{,}2904 - 0{,}2744}{0{,}2904} = 5{,}5\,\% \tag{28}$$

lo que confirma la tendencia observada en el máximo anterior.

En términos cualitativos, la tendencia global observada en el gráfico 3(b) parece indicar que inicialmente la medida de similitud tiene mayor capacidad para agrupar los resultados *fáciles* entre los primeros elementos del conjunto de recuperación. Cuando el conjunto de recuperación crece, la distancia euclídea incluye estos mismos resultados fáciles, aunque más alejados de los primeros lugares, por lo que las precisiones tienden a nivelarse, tendencia que se observa en la sección central del gráfico. Finalmente, la diferencia vuelve a aumentar, lo que indicaría que la medida de similitud tiene mejor capacidad para traer a los primeros lugares del conjunto de recuperación los casos más difíciles, que aparecen sólo después de haber aparecido varios falsos positivos.

Con objeto de desechar cualquier resultado casual en el experimento descrito, y confirmar definitivamente el mejor comportamiento de la medida de similitud propuesta en sistemas de evaluación de similitud, se repitió el experimento exactamente igual que en el caso que se acaba de describir, salvo por dos diferencias:

- el orden de los momentos de Zernike utilizado fue de 9 en lugar de 17
- el número de elementos recuperados para cada consulta fue 10, para simplificar la tarea de validación

Con estos nuevos parámetros, que suponen una representación más pobre para cada imagen, los resultados alcanzados son los que se muestran en la figura 4(a), donde se puede apreciar que se repite la tendencia decreciente en los valores de precisión media a medida que aumenta el conjunto recuperado.

Como puede apreciarse en la figura 4(b) la serie de diferencias es ahora prácticamente monótona creciente, indicando que con esta representación empobrecida la medida de similitud tiene mejor capacidad de recuperar imágenes similares para cualquier tamaño del conjunto de recuperación entre los explorados que la distancia euclídea. Para verificar si estadísticamente esta tendencia es significativa, se ha tomado como referencia el valor final de la serie, para conjuntos de recuperación de diez imágenes, y se ha repetido el test del signo. El resultado obtenido en esta ocasión para el estadístico ha sido de $\rho = 1{,}85\ 10^{-}7$, lo que indica que claramente la medida de similitud tiene mejor comportamiento que la distancia euclídea.

Al igual que en los casos anteriores, se ha cuantificado esta mejoría en el comportamiento en la evaluación de similitud, a partir de los valores de las precisiones medias:

- para la distancia euclídea es de $\Pi_D = 0{,}402$
- para la medida de similitud es de $\Pi_{MD} = 0{,}451$
- la mejora de precisión es por tanto de un $\Delta\Pi = 4{,}9\,\%$
- la mejora relativa de precisión es:

$$\Delta_{rel}\Pi = \frac{\Pi_{MD} - \Pi_D}{\text{máx}(\Pi_{MD}, \Pi_D)} = \frac{0{,}451 - 0{,}402}{0{,}451} = 10{,}86\,\% \tag{29}$$

Este resultado confirma plenamente los anteriores, e indica que la medida de similitud propuesta en este trabajo supera con creces el comportamiento de la distancia euclídea como herramienta para la recuperación de imágenes por contenido.

Todos los resultados para cada uno de estos conjuntos de recuperación con su data completa se resumen en la tabla 1.

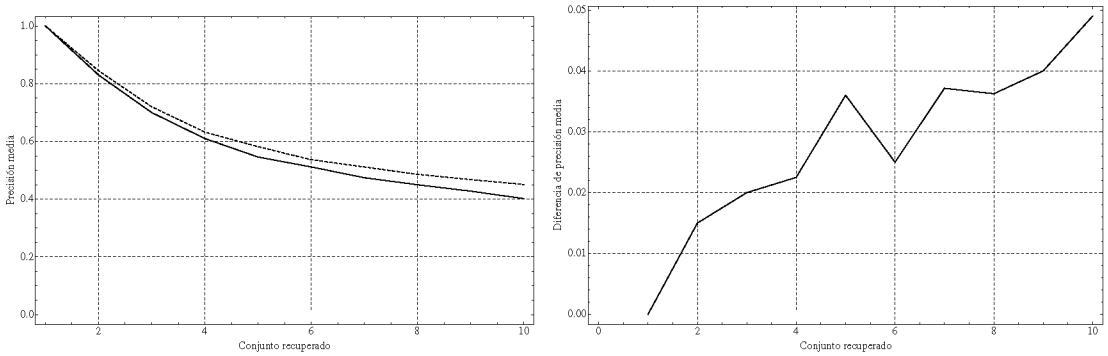
Tabla 1: Resumen de los resultados de los distintos expermientos

ω	M	Π_D	Π_{MD}	$\Delta\Pi$	$\Delta_{rel}\Pi$
17	5	0.57	0.6	3 %	**5 %**
17	25	0.2744	0.2904	1.6 %	**5.5** %
9	10	0.402	0.501	4.9 %	**10.86 %**

5. Conclusiones

En este trabajo se propone una medida de similitud basada en la extracción de características de una imagen y en su comparación siguiendo una estrategia de ponderación fundamentada en el concepto de saliencia. Se han sentado los principios formales que deben regir su definición a través de una serie de proposiciones, cuyo objetivo es cuantificar el efecto cualitativo de dicha saliencia para cada característica empleada, así como la estrategia a seguir en su combinación al resolver una tarea de comparación entre dos estímulos.

Los resultados obtenidos en diversos experimentos muestran un comportamiento nítidamente mejor que el referente usual en el desarrollo de las tareas de comparación en VIRS, como es la distancia euclídea aplicada a los vectores de características; la condición para alcanzar este resultado es que, dada una característica usada en la comparación, sus probabilidades de alcanzar distintos valores sean significativamente diferentes. Dicha mejoría se ha confirmado a través de los correspondientes contrastes de hipótesis, que confirman su relevancia estadística.



(a) Comportamiento neto de la precisión para distancia euclídea (continua) y medida de similitud (discontinua)

(b) Diferencia entre la precisión de la medida de similitud y de la distancia euclidea

Figura 4: Resultados de la validación del segundo expermiento.

Todas estas conclusiones indican la validez de la estrategia propuesta, e impulsan la profundización en el estudio de la definición de las funciones de ponderación de características; así, la formulación de nuevas formas para los factores de saliencia y de coincidencia, y más precisamente, la definición diferenciada las saliencias bottom-up y top-down, son líneas abiertas de investigación, de las que se esperan nuevos y mejores resultados.

English Summary

Saliency-based similarity measure

Abstract

The ubiquitous growth of multimedia production is causing the creation of new visual information retrieval paradigms. One of the most relevant among them is that represented by Visual Information Retrieval Systems (VIRS), where a common task is ordering a set images according to their similarity to a given one. In this work a new proposal for evaluating similarity between two images is introduced; both images are represented by respective feature vectors, and the perceptual cue used to generate the similarity measure is saliency, a concept thoroughly known in Psychology. New methodologies for quantifying saliency of feature values, for combining them during a comparison process and, eventually, to weight that feature attending to the result of the combination, are introduced as well. The results for the evaluation of this similarity measure in an image based content retrieval task are presented, as well as their comparison with those obtained using euclidean distance in the same task. Both are validated by volunteers who labelled the retrieved sets.

Keywords:

Image databases, content based retrieval, similarity measures, perceptual models, image analysis.

Referencias

Ashby, F., Perrin, N., 1988. Toward a unified theory of similarity and recognition. Psychological Review 95 (1), 124–150.

Chen, G., Xie, W., 2011. Wavelet-based moment invariants for pattern recognition. Optical Engineering 50 (7).

Eidenberger, H., 2006. Evaluation and analysis of similarity measures for content based visual information retrieval. ACM Multimedia Systems Journal 12 (2), 71–87.

Fairhall, A. L., Lewen, G. D., Bialek, W., de Ryiter van Steveninck, R. R., August 2001. Efficiency and ambiguity in an adaptive neural code. Nature 412 (6849), 787–792.

Fisher, R., 2011.
URL: `homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm`

Itti, L., Baldi, P., 2009. Byesian surprise attracts human attention. Vision Research 49, 1295–1306.

Itti, L., Koch, C., March 2001. Computational modelling of visual attention. Nature Reviews Neuroscience 2, 194–203.

Itti, L., Koch, C., Niebur, E., November 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11), 1254–1259.

Kim, H.-K., Kim, J.-D., Sim, D.-G., Oh, D.-I., 2000. A modified zernike moment shape descriptor invariant to translation, rotation and scale for similarity-based image retrieval. IEEE International Conference on Multimedia 1, 307–310.

Kim, Y., Kim, W., 1998. Content-based trademark retrieval using a visually salient feature. Image and Vision Computing 16, 931–939.

Koch, C., Ullman, S., 1985. Shifts in selective visual attention: Towards the underlying neural circuitry. Human Neurobiology 4, 219–227.

Larkey, L. B., Markman, A. B., 2005. Processes of similarity judgement. Cognitive Science 29, 1061–1076.

Rao, R. P. N., Ballard, D. H., January 1999. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience 2 (1), 79–87.

Santini, S., Jain, R., September 1999. Similarity measures. IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (9), 871–883.

Shepard, R. N., June 1962a. The analysis of proximities: Multidimensional scaling with an unknown distance function. i. Psycometrika 27 (2), 125–140.

Shepard, R. N., September 1962b. The analysis of proximities: Multidimensional scaling with an unknown distance function.ii. Psycometrika 27 (3), 219–246.

Teh, C.-H., Chin, R. T., !'July 1988. On image analysis by the methods of moments. IEEE Transactions on Pattern Analysis and Machine Intelligence 10 (4), 496–512.

Treue, S., 2003. Visual attention: the where, what, how and why of saliency. Current Opinion in Neurobiology 13, 428–432.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., Nuflo, F., 1995. Modeling visual attention via selective tuning. Artificial Intelligence 78, 507–545.

Tversky, A., July 1977. Features of similarity. Psychological Review 84 (4), 327–352.

Tversky, A., Gati, I., 1982. Simiarity, separability and the triangle equation. Psychological Review 89, 123–154.

Urdan, T. C., 2005. Statistics in plain english. Lawrence Erlbaum Associates.